On the Foundations of *Multinomial* Sequence Based Estimation*

B. John Oommen[1] and Sang-Woon Kim[2]

[1] *Chancellor's Professor*, School of Computer Science, Carleton University, Ottawa, Canada: K1S 5B6. e-mail address: oommen@scs.carleton.ca.
[2] Dept. of Computer Engineering, Myongji University, Yongin, 17058 South Korea. e-mail address: kimsw@mju.ac.kr

Abstract. This paper deals with the relatively new field of sequence-based estimation which involves utilizing both the information in the observations and in their sequence of appearance. Our intention is to obtain Maximum Likelihood estimates by "extracting" the information contained in the observations when perceived as a *sequence* rather than as a *set*. The results of [15] introduced the concepts of Sequence Based Estimation (SBE) for the Binomial distribution. This current paper generalizes these results for the multinomial "two-at-a-time" scenario. We invoke a novel phenomenon called "Occlusion" that can be described as follows: By "concealing" certain observations, we map the estimation problem onto a lower-dimensional binomial space. Once these occluded SBEs have been computed, we demonstrate how the overall Multinomial SBE (MSBE) can be obtained by mapping several lower-dimensional estimates onto the original higher-dimensional space. We formally prove and experimentally demonstrate the convergence of the corresponding estimates

Keywords: *Estimation using Sequential Information, Sequence Based Estimation, Estimation of multinomials, Fused Estimation Methods, Sequential Information.*

1 Introduction

Estimation is the central aspect associated with the training phase of classification and Machine Learning. Since the *sequence*-based paradigm for supervised

* The first author is a *Fellow: IEEE* and *Fellow: IAPR*. The work was done while he was visiting at Myongji University, Yongin, Korea. He also holds an *Adjunct Professorship* with the Department of Information and Communication Technology, University of Agder, Grimstad, Norway. The work was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada and a grant from the National Research Foundation of Korea. This work was also generously supported by the National Research Foundation of Korea funded by the Korean Government (NRF-2012R1A1A2041661).

learning that is explored in this paper is relatively new, we shall first motivate its perspective. Estimation methods generally fall into various categories, including the Maximum Likelihood Estimates (MLEs) and the Bayesian family of estimates [1,3,4,7,20] which are well-known for having good computational and statistical properties. Consider the strategy used for developing the MLE of the parameter of a distribution, $f_X(\theta)$, whose parameter to be estimated is θ. The input to the estimation process is the set of points/observations $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, whose elements are assumed to be generated independently and identically as per the distribution, $f_X(\theta)$. The process for computing the Maximum Likelihood (ML) estimate involves deriving the likelihood function, i.e., the likelihood of the distribution, $f_X(\theta)$, generating the sample points/observations $\mathcal{X}$ given θ, which is then maximized (by traditional optimization or calculus methods) to yield the estimate, $\widehat{\theta}_{MLE}$. The general characteristic sought for is that the estimate $\widehat{\theta}_{MLE}$ converges to the true (unknown) θ with probability one, or in a mean square sense. The Bayesian schemes work with a similar goal, except that rather than them using Likelihood functions, they compute the posterior distributions assuming that θ itself is a random variable with a known distributional form. Bayesian and ML estimates generally possess desirable convergence properties. Indeed, the theory of estimation has been studied for hundreds of years [1,3,10,17,18,19], and it has been the backbone for the learning (training) phase of statistical pattern recognition systems [4,7,9,20,21].

Traditionally, the ML and Bayesian estimation paradigms work within the model that the data, from which the parameters are to be estimated, is known, and that it is treated as a *set*. The position that we respectfully submit is that traditional ML and Bayesian methods ignore and discard[1] valuable *sequence*-based information. The goal of this paper is to "extract" and "utilize" the information contained in the observations when they are perceived *both as a set* and in *their sequence of appearance*. Put in a nutshell, this paper deals with the relatively new field of sequence-based estimation in which the goal is to estimate the parameters of a distribution by maximally "squeezing" out the *set*-based and *sequence*-based information latent in the observations.

The consequences of solving this problem are potentially many. Estimation, as researchers in almost all fields of science and engineering will agree, is a fundamental issue, in which the practitioner is given a set of observations involving the random variable, and his task is to estimate the parameters which govern the generation of these observations. Since, by definition, the problem involves random variables, decisions, predictions, regressions and classification related to the problem are, in some way, dependent on the practitioner obtaining reliable estimates of the parameters that characterize the underlying random variables.

More specifically, suppose that the user received $\mathcal{X}$ as a sequence of data points as in a typical real-life (or real-time) application such as those obtained in a data-mining domain involving sequences, or in data involving radio or television news items. The question that we have investigated is the following: "Is

[1] This information is, of course, traditionally used when we want to consider *dependence* information, as in the case of Markov models and n-gram statistics.

there any information in the fact that in $\mathcal{X}$, x_i specifically precedes x_{i+1}?". Or in a more general case, "Is there any information in the fact that in $\mathcal{X}$, the **sequence** $x_i x_{i+1} \ldots x_{i+j}$ occurs $n_{i,i+1,\ldots i+j}$ times?". Our position, which we proved in [15] for binomial random variables, is that even though $\mathcal{X}$ is generated by an i.i.d. process, there is information in these pieces of sequential data which can be "maximally" utilized to yield the so-called family of Sequence Based Estimators (SBEs). The problem was initially studied in [15], but only for the case of binomial random variables.

As far as we know, apart from the results in [15], there are no other reported results which utilize sequential information in obtaining such estimates. Also, as highlighted in [15], unlike the use of sequence information in syntactic pattern recognition, grammatical inference and in modeling channels using Hidden Markov Models (which involve estimating the bigram and n-gram probabilities of *dependent* streams of data [2,4,6]), in our case, we assume that the elements in the stream of data, $\mathcal{X}$, occur *independently*, and yet have information not utilized by traditional MLE schemes.

The contributions of this paper can be catalogued as follows:

1. This paper lists the first reported results for obtaining the MLEs of the parameters (i.e., the vector of probabilities responsible for the generation) of a multinomial distribution, when the data is processed both as a *set* of observations and as a *sequence* in which the samples occur in the set. These estimates are called the Multinomial Sequence Based Estimates (MSBEs).
2. The paper pioneers the concept of obtaining MSBEs by invoking the phenomenon of "Occlusion" in which certain observations are hidden or concealed to first yield binomial SBEs, and these are subsequently fused to yield the MSBE.
3. The paper contains the formal results[1] for the MSBE schemes when the sequence is processed in pairs. They have all been experimentally verified.

To the best of our knowledge, apart from our previous results of [15], all of these are novel to the field of estimation, learning and classification. Also, In the interest of space and brevity, the proofs of the theoretical results presented here are omitted. They are found in [16].

2 On Obtaining MSBEs Using Occluded SBEs

Before we proceed with the theoretical and experimental results, it is necessary for us to formalize the notation that will be used[2].

Notation 1: To be consistent, we introduce the following notation.

- X is a multinomially distributed random variable, obeying the distribution S.

[1] The paper lists numerous theorems whose proofs are found in [16]. The results for longer subsequences (i.e, three-at-a-time, four-at-a-time etc.) are also found in [16].

[2] We apologize for this cumbersome notation, but this is unavoidable considering the complexity of the problem and the ensuing analysis.

- $\mathcal{X} = \{x_1, x_2, \ldots, x_J\}$ is a realization of a sequence of occurrences of X, where each $x_i \in \mathcal{D}$.
- An index $a \in \mathcal{D}$ is said to be the unconstrained variable in any computation if all the other estimates $\{s_i\}$ are specified in terms of s_a, where $i \neq a$. It will soon be clear that in any computation there can only be *a single* unconstrained variable. The other variables are defined in terms of it.
- $\mathcal{X}^{ab} = \{x_1, x_2, \ldots, x_{N_{ab}}\}$ is called the *Occluded* sequence of $\mathcal{X}$ (with N_{ab} items) with respect to a and b, if it is obtained from $\mathcal{X}$ by deleting the occurrences of all the elements except a and b. Whenever we refer to the sequence $\mathcal{X}^{ab} = \{x_1, x_2, \ldots, x_{N_{ab}}\}$, we always imply that the first variable (in this case a) is the unconstrained variable.
- Let $< j_1 j_2 \ldots, j_k >$ be the subsequence[1] examined in the *Occluded* sequence $\mathcal{X}^{ab}$, where each $j_m, (1 \leq m \leq k)$, is either a or b. Then[2]:
  - The BSBE, for s_a obtained by examining in $\mathcal{X}^{ab}$ the subsequence $< j_1 j_2 \ldots, j_k >$ will be given by $\widehat{q}_a\Big|^{ab}_{<j_1 j_2 \ldots, j_k>}$, where, as before, the first variable (in this case a) is the unconstrained variable.
  - Similarly, the BSBE, for s_b obtained by examining in $\mathcal{X}^{ab}$ the subsequence $< j_1 j_2 \ldots, j_k >$ will be given by $\widehat{q}_b\Big|^{ab}_{<j_1 j_2 \ldots, j_k>}$, where the first variable (in this case a) is the unconstrained variable.
- Consider the sequence $\mathcal{X}$ in which the index a is the unconstrained variable. Let $< j_1 j_2 \ldots, j_k >$ be the subsequence examined in the sequence $\mathcal{X}$, where each $j_m, (1 \leq m \leq k)$, is either a or '$*$', where each '$*$' is the *same* variable, say $c \in (\mathcal{D} - \{a\})$. Then:
  - The MSBE for s_a (where a is the unconstrained variable) obtained by examining in $\mathcal{X}$ the sequence $< j_1 j_2 \ldots, j_k >$ will be given by $\widehat{s}_a\Big|^{a}_{<j_1 j_2 \ldots, j_k>}$ where each j_i that is not a is replaced by a '$*$', and where each '$*$' is the *same* variable, say $c \in (\mathcal{D} - \{a\})$.
  - For any constrained variable b, the MSBE for s_b obtained by examining in $\mathcal{X}$ the sequence $< j_1 j_2 \ldots, j_k >$ will be given by $\widehat{s}_b\Big|^{ab}_{<j_1 j_2 \ldots, j_k>}$, where a is the unconstrained variable.
- Trivially, for all a and b:

$$\textstyle\sum_{b \neq a} \widehat{s}_b\Big|^{ab}_{<j_1 j_2 \ldots, j_k>} = 1 \text{ - } \widehat{s}_a\Big|^{a}_{<j_1 j_2 \ldots, j_k>}. \qquad \square$$

Example of Notation 1: Let $\mathcal{D} = \{1, 2, 3, 4\}$, and $\mathcal{X} = 134211232341122$. Then, the *Occluded* sequence $\mathcal{X}^{12}$ is obtained by erasing from $\mathcal{X}$ all occurrences of 3 and 4, and has the form $\mathcal{X}^{12} = 1211221122$. Observe that N_{12} is 10. Then:

[1] For the present, we consider non-overlapping subsequences. We shall later extend this to overlapping sequences when we report the experimental results.

[2] The reader must take pains to differentiate between the q's and the s's, because the former refer to the BSBEs and the latter to the MSBEs.

– If 1 is the unconstrained variable, the BSBE of s_1 obtained by examining $\mathcal{X}^{12}$ for all occurrences of the sequence $< 121 >$ will be given by $\widehat{q}_1\Big|^{12}_{<121>}$.
– If 2 is the unconstrained variable, the BSBE of s_4 obtained by examining all occurrences of the sequence $< 224 >$ will be given by $\widehat{q}_4\Big|^{24}_{<224>}$.
– If in any specific computation, 4 is the unconstrained variable, the MSBE of s_4 obtained by examining all occurrences of the sequence $< **4 >$ will be given by $\widehat{s}_4\Big|^{4}_{<**4>}$, and will be obtained by normalizing using the quantities $\widehat{s}_1\Big|^{41}_{<114>}$, $\widehat{s}_2\Big|^{42}_{<224>}$ and $\widehat{s}_3\Big|^{43}_{<334>}$. □

2.1 The Fundamental Theorem of Fusing Occluded Estimates

Our first task is to formulate how we can compute the MSBEs by utilizing information gleaned by the *Binomial* SBEs (BSBEs) obtained from the set of occluded sequences. Consider an occluded sequence, $\mathcal{X}^{ab}$, extracted from the original sequence, $\mathcal{X}$, by removing all the variables except a and b. In the sequence being examined, we choose one variable, say a to be the unconstrained variable. We shall first attempt to obtain BSBEs of the relative proportions of s_a and s_b, the quantities to be estimated, from $\mathcal{X}^{ab}$. Thereafter, we utilize the set of these relative proportions to compute the MSBEs of all the variables. We formalize this in what we call the *Fundamental Theorem of Fusing Occluded Estimates.*

Theorem 1. *For every pair of indices, a and b, let $\mathcal{X}^{ab}$ be the Occluded sequence, extracted from the original sequence, $\mathcal{X}$, by removing all the variables except a and b. If we consider a to be the unconstrained variable, we define $q_a = \frac{s_a}{s_a+s_b}$ and $q_b = \frac{s_b}{s_a+s_b}$, where $q_a + q_b = 1$. Now let $\widehat{q}_a\Big|^{ab}_{\pi(a,b)} \neq 0$ and $\widehat{q}_b\Big|^{ab}_{\pi(a,b)} = 1 - \widehat{q}_a\Big|^{ab}_{\pi(a,b)}$ be the BSBEs of q_a and q_b respectively based on the occurrence*[1] *of any specific subsequence $\pi(a,b)$. Then, if c is a dummy variable*[2] *representing any of the variables, the MSBEs of s_a and s_b obtained by examining the occurrences*[3] *of $\pi(a,b)$ in every $\mathcal{X}^{ab}$ are:*

$$\widehat{s}_a\Big|^{a}_{\pi(a,b)} = \frac{1}{\sum_{\forall c}\rho_c}, \qquad \textit{and} \qquad \widehat{s}_b\Big|^{ab}_{\pi(a,b)} = \frac{\widehat{q}_b\Big|^{ab}_{\pi(a,b)}}{\sum_{\forall c}\rho_c}, \tag{1}$$

where $\rho_a = 1$ and $\forall c \neq a, \rho_c = \dfrac{\widehat{q}_c\Big|^{ac}_{\pi(a,c)}}{\widehat{q}_a\Big|^{ac}_{\pi(a,c)}}$.

[1] How BSBEs are obtained for specific instantiations of $\pi(a,b)$ is discussed later.
[2] The fact that c is a dummy variable will not be repeated in future invocations.
[3] This, of course, makes sense only if $\forall c, \widehat{q}_a\Big|^{ac}_{\pi(a,c)} \neq 0$.

Proof. This is the central theorem of this paper. With a being unconstrained, let the BSBE of q_a based on the occurrence of any specific subsequence $\pi(a,b)$ be $\widehat{q}_a\Big|_{\pi(a,b)}^{ab}$. Clearly, $\widehat{q}_b\Big|_{\pi(a,b)}^{ab} = 1 - \widehat{q}_a\Big|_{\pi(a,b)}^{ab}$. The MSBE is then obtained by resorting to the Weak Law of Large Numbers which guarantees that if the sequence examined is "large enough", the ratio between the various probabilities is also the ratio between their estimates, thus providing a mechanism to normalize the corresponding estimates.

The proof of the result is omitted due to space considerations. It is in [16]. □

3 MSBEs Using Pair-wise Sequential Information

3.1 Theoretical Results

The following results for MSBEs are true when the sequential information is processed in pairs.

Theorem 2. *Let $q_a = \frac{s_a}{s_a+s_b}$ and $q_b = \frac{s_b}{s_a+s_b}$, where $q_a + q_b = 1$. Then, $\widehat{q}_a\Big|_{<aa>}^{ab}$ and $\widehat{q}_b\Big|_{<aa>}^{ab}$, the BSBEs of q_a and q_b obtained by examining the occurrences of $< aa >$ in $\mathcal{X}^{ab}$ are:*

$$\widehat{q}_a\Big|_{<aa>}^{ab} = \sqrt{\frac{n_{aa}}{N_{ab}/2}}, \qquad \text{and} \qquad \widehat{q}_b\Big|_{<aa>}^{ab} = 1 - \sqrt{\frac{n_{aa}}{N_{ab}/2}}, \tag{2}$$

where n_{aa} is the number of occurrences of $< aa >$ from among the $\frac{N_{ab}}{2}$ non-overlapping subsequences[1] of length 2 in $\mathcal{X}^{ab}$. Consequently,

$$\widehat{s}_a\Big|_{<aa>}^{a} = \frac{1}{\sum_{\forall c}\rho_c}, \qquad \text{and} \qquad \widehat{s}_b\Big|_{<aa>}^{ab} = \frac{\widehat{q}_b\Big|_{<aa>}^{ab}}{\sum_{\forall c}\rho_c}, \tag{3}$$

where $\rho_a = 1$ and $\forall c \neq a, \rho_c = \frac{1-\sqrt{\frac{n_{aa}}{N_{ac}/2}}}{\sqrt{\frac{n_{aa}}{N_{ac}/2}}}$.

Proof. The proof of the theorem is found in [16]. □

The following example will help clarify the concepts of how the BSBEs are computed and how the MSBE is obtained from the BSBEs.

Example I: Let us suppose that:

$\mathcal{X} = \{2,2,3,3,1,1,2,1,1,2,3,2,3,1,1,2,1,1,2,2,2,1,3\}$.

We shall consider the MSBEs for the case when the variable 1 is unconstrained. This will highlight why our present results are *far more complex* than the corresponding binomial results derived in [15]. Indeed, the extension of the binomial to the multinomial case depends *on the identity of the unconstrained variable.*

[1] Observe that it would be statistically advantageous (since the number of occurrences obtained would be almost doubled) if all the overlapping $N_{ab} - 1$ subsequences of length 2 were considered. The computational consequences of this are given in [16].

Estimation of the MSBE when 1 is the Unconstrained Variable

First of all, $\mathcal{X}^{12} = \{2,2,1,1,2,1,1,2,2,1,1,2,1,1,2,2,2,1\}$,
and, $\mathcal{X}^{13} = \{3,3,1,1,1,1,3,3,1,1,1,1,1,3\}$.

From the set $\mathcal{X}^{12}$, we see that $N_{12} = 18$, and in $\mathcal{X}^{12}$, $n_{11} = 4$, and so, as per Eq. (2):

$$\widehat{q}_1\Big|^{12}_{<11>} = \sqrt{\tfrac{4}{9}} = \tfrac{2}{3}, \text{ and}$$

$$\widehat{q}_2\Big|^{12}_{<11>} = 1 - \sqrt{\tfrac{4}{9}} = \tfrac{1}{3}.$$

Thus, $\widehat{q}_2\Big|^{12}_{<11>} = (\mathbf{0.5}) \cdot \widehat{q}_1\Big|^{12}_{<11>}$.

Again, from the set $\mathcal{X}^{13}$, we see that $N_{13} = 14$, and in $\mathcal{X}^{13}$, $n_{11} = 4$, and so, as per Eq. (2):

$$\widehat{q}_1\Big|^{14}_{<11>} = \sqrt{\tfrac{4}{7}} = 0.7559, \text{ and}$$

$$\widehat{q}_3\Big|^{13}_{<11>} = 1 - \sqrt{\tfrac{4}{6}} = 0.2441.$$

Thus, $\widehat{q}_3\Big|^{13}_{<11>} = (\mathbf{0.323}) \cdot \widehat{q}_1\Big|^{13}_{<11>}$.

Normalizing the above with regard to the relative proportions to variable 1 as per Theorem 1, implies normalizing $[\theta \quad 0.5\theta \quad 0.323\theta]^T$. This yields the MSBE of $[s_1 \quad s_2 \quad s_3]^T$, with 1 being the unconstrained variable, to be $[0.5485 \quad 0.2743 \quad 0.1772]^T$. □

The corresponding results for $\widehat{s}_a\Big|^{a}_{<bb>}$, $\widehat{s}_a\Big|^{a}_{<ab>}$ and $\widehat{s}_a\Big|^{a}_{<ba>}$ etc. follow.

Theorem 3. *Let $q_a = \frac{s_a}{s_a+s_b}$ and $q_b = \frac{s_b}{s_a+s_b}$, where $q_a + q_b = 1$. Then, $\widehat{q}_a\Big|^{ab}_{<bb>}$ and $\widehat{q}_b\Big|^{ab}_{<bb>}$, the BSBEs of q_a and q_b obtained by examining the occurrences of $< bb >$ in $\mathcal{X}^{ab}$ are:*

$$\widehat{q}_a\Big|^{ab}_{<bb>} = 1 - \sqrt{\frac{n_{bb}}{N_{ab}/2}}, \qquad \textit{and} \qquad \widehat{q}_b\Big|^{ab}_{<bb>} = \sqrt{\frac{n_{bb}}{N_{ab}/2}}. \tag{4}$$

where n_{bb} is the number of occurrences of $< bb >$ from among the $\frac{N_{ab}}{2}$ non-overlapping subsequences of length 2 in $\mathcal{X}^{ab}$. Consequently,

$$\widehat{s}_a\Big|^{a}_{<bb>} = \frac{1}{\sum_{\forall c} \rho_c}, \qquad \textit{and} \qquad \widehat{s}_b\Big|^{ab}_{<bb>} = \frac{\widehat{q}_b\Big|^{ab}_{<bb>}}{\sum_{\forall c} \rho_c}, \tag{5}$$

where $\rho_a = 1$ and $\forall c \neq a, \rho_c = \frac{\sqrt{\frac{n_{cc}}{N_{ac}/2}}}{1-\sqrt{\frac{n_{cc}}{N_{ac}/2}}}$.

Proof. The proof is similar to that of Theorem 2. The details are omitted. □

Theorem 4. *Let* $q_a = \frac{s_a}{s_a+s_b}$ *and* $q_b = \frac{s_b}{s_a+s_b}$, *where* $q_a + q_b = 1$. *Then,* $\widehat{q}_a\Big|^{ab}_{<ab>}$, *the BSBE of* q_a *obtained by examining the occurrences of* $< ab >$ *in* $\mathcal{X}^{ab}$, *can be obtained if and only if the roots of the quadratic equation* $\lambda^2 - \lambda + \frac{n_{ab}}{N_{ab}/2} = 0$ *are real (where* n_{ab} *is the number of occurrences of* $< ab >$ *from among the* $\frac{N_{ab}}{2}$ *non-overlapping subsequences of length* 2 *in* $\mathcal{X}^{ab}$*). Its value,* λ_a, *is the root whose value is closest to* $\widehat{q_a}$. *Further, in such a case,* $\widehat{q}_b\Big|^{ab}_{<ab>} = 1 - \widehat{q}_a\Big|^{ab}_{<ab>} = \lambda_b$. *Finally,*

$$\widehat{s}_a\Big|^{a}_{<ab>} = \frac{1}{\sum_{\forall c}\rho_c}, \qquad \textit{and} \qquad \widehat{s}_b\Big|^{ab}_{<ab>} = \frac{\widehat{q}_b\Big|^{ab}_{<ab>}}{\sum_{\forall c}\rho_c}, \tag{6}$$

where $\rho_a = 1$ *and* $\forall c \neq a, \rho_c = \frac{\lambda_c}{\lambda_a}$.

Proof. The proof of this theorem is also included in [16]. □

The final theorem about the MSBE computed using the occurrences of $< ba >$ in $\mathcal{X}^{ab}$ is given below. Its proof is identical to the one above.

Theorem 5. *Let* $q_a = \frac{s_a}{s_a+s_b}$ *and* $q_b = \frac{s_b}{s_a+s_b}$, *where* $q_a + q_b = 1$. *Then,* $\widehat{q}_a\Big|^{ab}_{<ba>}$, *the BSBEs of* q_a *obtained by examining the occurrences of* $< ba >$ *in* $\mathcal{X}^{ab}$, *can be obtained if and only if the roots of the quadratic equation* $\lambda^2 - \lambda + \frac{n_{ba}}{N_{ab}/2} = 0$ *are real (where* n_{ba} *is the number of occurrences of* $< ba >$ *from among the* $\frac{N_{ab}}{2}$ *non-overlapping subsequences of length* 2 *in* $\mathcal{X}^{ab}$*). Its value,* λ_a, *is the root whose value is closest to* $\widehat{q_a}$. *Further, in such a case,* $\widehat{q}_b\Big|^{ab}_{<ba>} = 1 - \widehat{q}_a\Big|^{ab}_{<ba>} = \lambda_b$. *Finally,*

$$\widehat{s}_a\Big|^{a}_{<ba>} = \frac{1}{\sum_{\forall c}\rho_c}, \qquad \textit{and} \qquad \widehat{s}_b\Big|^{ab}_{<ba>} = \frac{\widehat{q}_b\Big|^{ab}_{<ba>}}{\sum_{\forall c}\rho_c}, \tag{7}$$

where $\rho_a = 1$ *and* $\forall c \neq a, \rho_c = \frac{\lambda_c}{\lambda_a}$. □

Notice that the four estimates $\widehat{s}_a\Big|^{a}_{<aa>}$, $\widehat{s}_a\Big|^{a}_{<ab>}$, $\widehat{s}_a\Big|^{a}_{<ba>}$ and $\widehat{s}_a\Big|^{a}_{<bb>}$ are not linearly independent. Indeed, this is true because: $n_{aa} + n_{ab} + n_{ba} + n_{bb} = \frac{N_{ab}}{2}$.

3.2 Experimental Results: Sequences of Pairs

In this section, we present the results of our simulations[1] on synthetic data for the case when the sequence is processed in pairs. In every case, we have considered the $N_{ab} - 1$ *overlapping* subsequences of length 2 for the occluded sequence $\mathcal{X}^{ab}$.Thus, for all $b \neq a$, we have used the following expressions to obtain *computational* approximations of the true corresponding estimates derived in Theorems 2 to 5 respectively:

[1] In the tables, values of *unity/zero* represent the cases when the roots are complex or when the number of occurrences of the event concerned are zero.

$\widehat{q}_a\Big|_{<aa>}^{ab} = \sqrt{\frac{n_{aa}}{N_{ab}-1}}$,
The roots of $\lambda^2 - \lambda + \frac{n_{ab}}{N_{ab}-1} = 0$,
The roots of $\lambda^2 - \lambda + \frac{n_{ba}}{N_{ab}-1} = 0$, and
$\widehat{q}_a\Big|_{<bb>}^{ab} = 1 - \sqrt{\frac{n_{bb}}{N_{ab}-1}}$,
where in each case, we have used $(N_{ab} - 1)$ instead of $\frac{N_{ab}}{2}$.

In every case examined, the multinomial distribution was S, where $S = [s_1, s_2 \dots s_d]^T$, with d taking values 3, 4 and 5.

The MSBE process for the estimation of S was extensively tested for numerous distributions and for different dimensionalities, but in the interest of brevity, we merely cite a single specific example for a given value of d. In each case, the estimation algorithms were presented with random occurrences of the variables for $N = 390625$ (i.e, 5^8) time instances. Each table reports the results of the estimation for the specific value of d, and in each table, the respective actual value of S used has been specified. To render the comparison meaningful, we have also used the identical data stream to follow the "traditional" MLE computation, i.e., the one that does not utilize the sequential information.

To compare the value of S to its estimate, we have also computed the Euclidean distance between S and its estimates, $\hat{S}$, namely $E_{MLE} = ||S - \widehat{S_{MLE}}||$ and $E_{MSBE} = ||S - \widehat{S_{MSBE}}||$, where $\widehat{S_{MLE}}$ was the ML estimate, and $\widehat{S_{MSBE}}$ was evaluated using the corresponding result depending on the pair of symbols examined in the occluded sequence. The results are tabulated in [16] and respectively, and when the pairs examined in every $\mathcal{X}^{ab}$ were aa, ab, ba and bb. To demonstrate the true convergence properties of the estimates, we have also reported the values of the ensemble averages of the estimates in Tables 1 and 2 respectively, taken over an ensemble of 100 experiments. The convergence of every single estimate is remarkable.

To be more specific, for the case when $d = 3$ and $S = [0.6 \quad 0.25 \quad 0.15]^T$ and when the pair examined in every $\mathcal{X}^{ab}$ was aa, the E_{MSBE} had the ensemble average of 0.1263 when only $N = 25$ symbols were processed (please see Table 1). This value decreased to 0.1247 when $N = 125$ symbols were processed. This error was marginally lower (due to the sampling variance) than the asymptotic error at $N = 5^8$ of 0.1272. The reader should also observe the manner in which the E_{MSBE} closely followed the E_{MLE}.

By way of comparison, when the pair examined in every $\mathcal{X}^{ab}$ was ab, (again for the case when $d = 3$) the value E_{MSBE} had the ensemble average of 0.1740 when only $N = 25$ symbols were processed. The progressive decrease of the error was again observed. It became 0.1412 when $N = 125$ symbols were processed, and became very close to the steady-state value when even as few as 625 samples were examined. Due to the sampling error caused by the random sequences, the MLE and MSBEs taken for a *single* experiment don't follow such a regular pattern, especially for small values of N.

Due to space limitations, the theoretical and experimental results for the cases when the subsequences are of lengths 3 and 4 are found in [16].

Table 1. A table of the *ensemble* averages (taken over 100 experiments) of E_{MLE}, the error of the MLE, and the error of the MSBE, E_{MSBE}, at time N, for $d = 3$, where the latter MSBEs were estimated by using the formal expressions of Theorems 2 to 5 approximated using the issues discussed in the beginning of this section. Here $d = 3$ and $S = [0.6 \quad 0.25 \quad 0.15]^T$. In the case of the MSBE, in each column, we mention the pair being examined, i.e., whether it is $< aa >$, $< ab >$, $< ba >$ or $< bb >$.

N	E_{MLE}	$E_{MSBE}\mid_{<aa>}$	$E_{MSBE}\mid_{<ab>}$	$E_{MSBE}\mid_{<ba>}$	$E_{MSBE}\mid_{<bb>}$
5^2 (25)	0.1091	0.1263	0.1740	0.1712	0.1725
5^3 (125)	0.1221	0.1247	0.1412	0.1036	0.1122
5^4 (625)	0.1252	0.1258	0.1292	0.1248	0.1269
5^5 (3,125)	0.1270	0.1272	0.1277	0.1284	0.1281
5^6 (15,625)	0.1273	0.1274	0.1273	0.1281	0.1280
5^7 (78,125)	0.1272	0.1272	0.1270	0.1274	0.1273
5^8 (390,625)	0.1272	0.1272	0.1272	0.1273	0.1273

Table 2. A table of the *ensemble* averages (taken over 100 experiments) of E_{MLE}, the error of the MLE, and the error of the MSBE, E_{MSBE}, at time N, for $d = 5$, where the latter MSBEs were estimated by using the formal expressions of Theorems 2 to 5 approximated using the issues discussed in the beginning of this section Here $d = 5$ and $S = [0.33 \quad 0.25 \quad 0.18 \quad 0.14 \quad 0.10]^T$. In the case of the MSBE, in each column, we mention the pair being examined.

N	E_{MLE}	$E_{MSBE}\mid_{<aa>}$	$E_{MSBE}\mid_{<ab>}$	$E_{MSBE}\mid_{<ba>}$	$E_{MSBE}\mid_{<bb>}$
5^2 (25)	0.1363	NaN	0.2217	0.2242	0.2120
5^3 (125)	0.1696	0.1746	0.1952	0.1751	0.1581
5^4 (625)	0.1864	0.1861	0.1932	0.1516	0.1541
5^5 (3,125)	0.1862	0.1856	0.1906	0.1466	0.1468
5^6 (15,625)	0.1882	0.1883	0.1888	0.1495	0.1497
5^7 (78,125)	0.1879	0.1879	0.1881	0.1769	0.1770
5^8 (390,625)	0.1880	0.1879	0.1880	0.1882	0.1882

4 Conclusions

In this paper, we have investigated the relatively new field of sequence-based estimation. The pioneering work in this area [15] introduced the concepts of Sequence Based Estimation (SBE) for Binomial distributions. This paper has generalized the latter results for multinomial distributions. The rationale motivating the development of SBEs and MSBEs is that traditional ML and Bayesian estimation ignore/discard valuable *sequence*-based information. SBEs "extract" the information contained in the observations when perceived as a *sequence*. In this paper, we have generalized the results of [15] for the multinomial case. Our strategy involves a novel and previously-unreported phenomenon called "Occlusion" where by hiding (or concealing) certain observations, we map the original estimation problem onto a lower-dimensional binomial space. We have also shown

how these consequent occluded SBEs can be fused to yield overall Multinomial SBE (MSBE). This is achieved by mapping several lower-dimensional estimates, that are all bound by rigid probability constraints, onto the original higher-dimensional space. The theoretical and experimental results for the cases when the subsequences are of lengths 3 and 4 are found in [16].

References

1. P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume I. Prentice Hall, Second Edition, 2000.
2. H. Bunke. Structural and Syntactic Pattern Recognition. *Handbook of Pattern Recognition and Computer Vision*, World Scientific-25, 1993.
3. G. Casella and R. Berger. *Statistical Inference*. Brooks/Cole Pub. Co., Second Edition, 2001.
4. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, NY, Second Edition, 2000.
5. M. A. El-Gendy, A. Bose, and K. G. Shin, "Evolution of the internet QoS and support for soft real-time applications," *Proceedings of the IEEE*, vol. 91, pp. 1086–1104, Jul. 2003.
6. M. Friedman and A. Kandel, *Introduction to Pattern Recognition - Statistical, Structural, Neural and Fuzzy Logic Approaches*, World Scientific, New Jersey, 1999.
7. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
8. S. Goldberg, *Probability: An Introduction*, Prentice-Hall, Englewood Cliffs, New Jersey, 1960.
9. R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Cambridge, Massachusetts, 2001.
10. B. Jones, P. Garthwaite, and Ian Jolliffe. *Statistical Inference*. Oxford University Press, Second Edition, 2002.
11. J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, "On combining classifiers", *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. PAMI-20, pp. 226 - 239, Mar. 1998.
12. E. Kreyszig, *Advanced Engineering Mathematics, Eighth Edition*, John Wiley & Sons, New York, 1999.
13. L. I. Kuncheva, J. C. Bezdek and R. P. W. Duin "Decision templates for multiple classifier fusion: an experimental comparison", *Pattern Recognition*, Vol. 34, pp. 299 - 414, Feb. 2001.
14. L. I. Kuncheva, "A theoretical study on six classifier fusion strategies", *IEEE Trans. Pattern Anal. and Machine Intell.*, Vol. PAMI-24, pp. 281 - 286, Feb. 2002.
15. B. J.Oommen, S-W. Kim and G. Horn, "On the Estimation of Independent Binomial Random Variables Using Occurrence and Sequential Information", *Pattern Recognition*, pp. 3263-3276, November 2007.
16. B. J.Oommen and S-W. Kim, "Occlusion-based Estimation of Independent *Multinomial* Random Variables Using Occurrence *and* Sequential Information". To be submitted for Publication.
17. S. Ross. *Introduction to Probability Models*. Academic Press, second edition, 2002.
18. J. Shao. *Mathematical Statistics*. Springer Verlag, second edition, 2003.
19. R. Sprinthall. *Basic Statistical Analysis*. Allyn and Bacon, second edition, 2002.
20. F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, parameter estimation and state estimation: An Engineering Approach using MATLAB*, John Wiley and Sons, Ltd., England, 2004.
21. A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, N.York, Second Edition, 2002.